反向散射辅助的无线供能通信中的信息年龄最小化 *

宋兆希，唐　冬†，黄高飞，赵　赛，刘贵云

(广州大学 电子与通信工程学院，广州 510006)

摘　要：信息年龄(AoI)是一种从目的端的角度衡量所捕获数据新鲜度的性能指标。在能量受限的实时感知物联网场景中，为了提高系统的 AoI 性能，提出了联合采样和混合反向散射通信更新的策略。该策略通过允许源端选择状态采样动作以及更新过程的传输模式来最小化系统的长期平均 AoI。具体来说，首先将该优化问题建模为一个平均成本马尔可夫决策过程(MDP)，然后在已知环境动态信息的情况下，通过相关值迭代算法获取最优策略；在缺乏环境动态信息的情况下，采用 Q 学习算法和探索利用方法，通过与环境的试错交互来学习最优策略。仿真结果表明，与两种参考策略相比，所提出的策略明显提高了系统 AoI 性能，同时发现系统的 AoI 性能随更新包尺寸的减小或者电池容量的增大而提升。

关键词：信息年龄；无线供能通信；反向散射通信；马尔可夫决策过程；强化学习；Q 学习

中图分类号：TP393;　　doi: 10.19734/j.issn.1001-3695.2021.12.0691

Age of information minimization for backscatter assisted wireless powered communications

Song Zhaoxi, Tang Dong†, Huang Gaofei, Zhao Sai, Liu Guiyun

(*School of Electronics & Communication Engineering*, *Guangzhou University*, *Guangzhou* 510006, *China*)

Abstract: Age of Information (AoI) is a performance metric that captures the freshness of data from the destination's perspective. In the energy constrained real-time sensing Internet of things scenario, this paper proposed a joint sampling and hybrid backscatter communication updating policy to improve the AoI performance of the system. The policy minimized the long-term average AoI of the system by allowing the source to select state sampling actions and transmission modes of updating processes. Specifically, this paper modeled the problem as an average cost Markov decision process (MDP) . Then, when the system realized the dynamic environment information, the paper adopted optimal strategy by relative value iterative algorithm. When the system lacked the dynamic environment information, the paper applied Q-learning algorithm and exploration exploitation technique to learn the optimal strategy through trial-and-error interactions with the environment. Simulation results show that compared with the two reference policies, the proposed policy significantly improves the AoI performance of the system, and the AoI performance of the system increases with the decrease of the update package size or the increase of battery capacity.

Key words: age of information; wireless powered communication; backscatter communication; Markov decision process; reinforcement learning; Q-learning

0　引言

随着物联网技术的发展，近年来越来越多的无线传感器节点被部署到各种实时状态监控系统中，例如环境监测、智能交通和智能农业系统等等。这些物联网应用基于对物理过程的实时状态更新来输出决策，决策的准确性取决于接收信息的新鲜程度[1]。为了衡量和量化接收信息的新鲜程度，文献[2]提出了信息年龄(age of information, AoI)，它从目的端的角度对接收信息的新鲜程度进行量化，定义为自源端生成的最新状态更新成功到达目的端所经过的时间，时间越短(AoI 值越小)新鲜度越好(AoI 性能越好)。然而，物联网设备的能量受限特性导致设备无法及时地发送更新，从而增加了物联网应用收到过时状态更新的可能性。能量收集(energy harvesting, EH)技术被认为是最有希望解决这一问题的方案之一，它的发展大大缓解了物联网设备能量受限的问题。它可以通过捕获周围的动能、热能、太阳能或者射频能量(radio frequency, RF)并转换为电能来保持设备的持久运行[3,4]。特别是由于无线电波的无处不在，基于射频的无线能量传输(wireless power transfer, WPT)被认为是有潜力的一种能量收集技术。另一方面，由于反向散射通信(backscatter communication, BC)技术具有超低功耗的特点，可广泛应用于能量受限的物联网和无线传感器网络场景中，以降低设备的通信能耗和运行成本。因此，在时间敏感的物联网网络中考虑结合 WPT 技术和 BC 技术可以减小系统的整体能耗，实现在维持网络设备监测服务持续运行的同时保持物联网应用接收信息的新鲜度。

AoI 的早期工作主要集中在从排队论的角度最小化 AoI，即通过将更新系统建模为由源、服务设施、监视器组成的队列系统，并利用最优化理论工具来最小化 AoI[2,5]。最近，文献[6~8]研究了在能量收集通信系统中 AoI 的分析和优化，其中源端使用从自然界中获取的能量进行更新传输，并且由于能量产生的不可预测性，能量收集过程通常被建模为独立的随机过程。然而，当源端从周围的射频信号中进行能量收集

收稿日期：2021-12-23；**修回日期：**2022-02-16　　**基金项目：**国家自然科学基金资助项目(61902084，61872098)；广东省教育厅广东高校特色创新项目(2018KTSCX174)

作者简介：宋兆希(1997-)，男，广东梅州人，硕士研究生，主要研究方向为反向散射通信；唐冬(1967-)，男(通信作者)，教授，博士，主要研究方向为新一代移动与无线通信理论等(tangdong@gzhu.edu.cn)；黄高飞(1978-)，男，副教授，博士，主要研究方向为无线信息与能量同传、移动边缘计算、无人机通信等；赵赛(1981-)，女，副教授，博士，主要研究方向为新一代无线通信关键技术、数据驱动无线通信等；刘贵云(1983-)，男，副教授，博士，主要研究方向为无线网络安全.

时[9~11]，收集的能量大小将依赖于射频源的发射功率和当前时隙的信道状态信息(channel state information, CSI)。文献[12]进一步考虑了更新的生成时间并提出了一种联合采样和更新策略。在该策略中，源端需要决定更新包的生成和发送时间，然后在需要发送时通过无线供能通信(wireless powered communication, WPC)实现状态更新包的传输。然而，由于 WPC 需要消耗大量的能量进行主动信息传输，这导致了高功耗问题，进一步加剧了源端的电池能量限制。

不同于 WPC，BC 是一种新兴的绿色低功耗通信技术[13]，它是实现可持续通信的一种有希望的选择。具体地，BC 可以通过反射来自外部射频源的入射信号来进行信息传输，它不需要产生主动射频信号，所以消耗的功率要比 WPC 低几个数量级。然而，BC 的传输范围有限且数据速率相对较低。为了克服 BC 的局限，文献[14~17]研究了一种结合 BC 和 WPC 的混合反向散射通信(hybrid backscatter communication, HBC)方案以最大化系统吞吐量性能，其中发射器可以自适应地选择 BC 或 WPC 进行数据传输。特别是文献[17]提出了一种新的混合通信协议，在该协议中混合发射器被允许以细粒度的方式在一个时间块内自适应地切换 EH、BC 或 IT 模式来进一步提高系统的吞吐量性能。然而，文献[14~17]并没有考虑到如何在反向散射辅助的无线供能通信中最小化系统的 AoI 值。

尽管在反向散射通信的研究中以 AoI 为性能指标的文献较少，但它依然是一个关键因素。因此，在时间敏感的物联网应用中，开发一种以最小化系统平均 AoI 为目标的采样和更新策略是本文的研究重点。 虽然文献[12]所提出的联合采样和 WPC 更新策略在一定程度上提高了系统的 AoI 性能，但是 WPC 的高功耗特性间接地限制了系统 AoI 性能的提高。在这种情况下，本文考虑结合 WPT 和 BC 技术实现状态更新的传输，通过运用基于模型的相关值迭代算法和无模型的 Q 学习算法[18]求解优化问题，提出了一种最小化系统长期平均 AoI 的联合采样和 HBC 更新策略，该策略通过允许源端根据当前信道状态、电池能量状态以及源端和目的端 AoI 信息自适应地选择状态采样动作和更新传输模式来进一步提高系统的 AoI 性能。

1　系统模型

系统模型如图 1 所示，考虑由一个能量发射器 ET、源端 S 和目的端 D 组成的无线反向散射传感器网络。其中，能量发射器 ET 连接到电网，用于向源端提供射频能量。源端包括一个能对物理过程进行实时状态采样的传感器和一个能向目的地发送状态更新信息的混合发射器。混合发射器配备射频能量收集电路、反向散射电路和主动射频电路，以便通过混合反向散射和无线供能通信实现射频能量的收集和状态信息的传输。

假设系统时间被划分为具有索引 $n = 0,1,2,\dots,N$ 的单位时隙。不失一般性，假设每个时隙的持续时间为 1 秒。源端 S 将在每个时隙的开始时刻决定采样动作和更新模式，并且状态采样和更新传输可以在一个时隙内完成。此外，本文考虑源端可以执行复杂的任务，因此状态采样的时间成本和能量成本不可忽略[19]。令 $h(n)$ 和 $g(n)$ 分别表示 n 时隙 ET 到 S、S 到 D 的信道链路增益，假设它们都受到准静态信道衰落的影响，这意味着信道状态将在一个时隙内保持不变，在不同时隙之间独立变化。

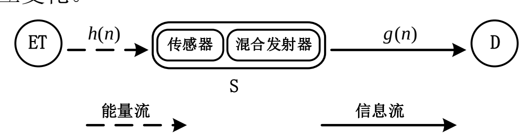


图 1　无线反向散射传感器网络模型

Fig. 1　Wireless backscatter sensor network model

1.1　监测模型

考虑一种联合采样和混合反向散射通信更新策略，即在 n 时隙的开始时刻，源端不仅需要决定传感器的状态采样动作，还需要决定混合发射器的状态更新模式。状态更新模式示意图如图 2 所示，在时隙 n 内，源端可以通过控制其内的混合发射器执行 EH 模式进行能量收集或者执行 BC、IT 等单一模式或者执行 EH-BC、EH-IT、BC-IT、EH-BC-IT 等组合模式进行状态更新的传输。特别地，为了易于处理，可以将 EH 模式表示为 a 模式，用于状态更新传输的单一模式 BC，IT 表示为 b 模式和 c 模式，并且组合模式 EH-BC、EH-IT、BC-IT、EH-BC-IT 分别对应表示为 d 模式、e 模式、f 模式以及 g 模式。

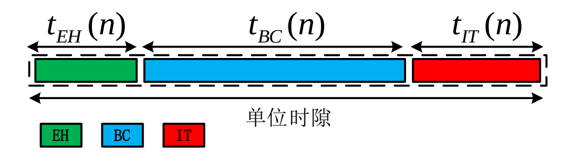


图 2　状态更新模式示意图

Fig. 2　Schematic diagram of state update mode

令 $a(n) = (w(n), z_m(n))$ 表示 n 时隙源端的状态采样和更新模式向量，其中 $w(n) \in \{0,1\}$ 为源端的状态采样动作， $z_m(n) \in \{0,1\}$ ，$m \in \mathcal{M} \triangleq \{a,b,c,d,e,f,g\}$ 表示源端的状态更新模式。若源端在 n 时隙进行状态采样则 $w(n)=1$，否则， $w(n)=0$ 。 $z_a(n)=1$ 表示源端在 n 时隙进行能量收集，否则，$z_k(n)=1$， $k \in \mathcal{M}' \triangleq \{b,c,d,e,f,g\}$ 表示源端在 n 时隙通过 k 模式传输状态更新。

1.2　能量收集模型

假设能量发射器 ET 以恒定功率 P 向源端 S 持续地发送射频能量。同时，源端将收集到的能量存储在容量为 $B_{\max}$ 的电池中，用于在未来进行状态信息的采样和更新包的传输。令 $\boldsymbol{t}(n) \in \{t_{EH}(n), t_{BC}(n), t_{IT}(n)\}$ 表示模式运行时间向量，其中 $t_{EH}(n)$ ，$t_{BC}(n)$ ，$t_{IT}(n)$ 它们分别表示时隙 n 中 EH，BC，IT 模式的运行时间。因此，对源端的不同模式，时间分配应满足如下约束:对于模式 a ，$t_{EH}(n)=1$ ，$t_{BC}(n)=t_{IT}(n)=0$ ；对于模式 b ，$t_{BC}(n)=1$ ，$t_{EH}(n)=t_{IT}(n)=0$ ；对于模式 c ，$t_{IT}(n)=1$，$t_{EH}(n)=t_{BC}(n)=0$ ；类似地，模式 d 有：$t_{IT}(n)=0$ ，$t_{EH}(n)+t_{BC}(n)=1$ ；模式 e 有：$t_{BC}(n)=0$ ，$t_{EH}(n)+t_{IT}(n)=1$ ；模式 f 有：$t_{EH}(n)=0$ ，$t_{BC}(n)+t_{IT}(n)=1$ ；最后，对于模式 g ，$t_{EH}(n)+t_{BC}(n)+t_{IT}(n)=1$ 。为了易于处理，上述等式可以表示为

$$
\begin{aligned}
& z_a(n)t_{EH}(n) + z_b(n)t_{BC}(n) + z_c(n)t_{IT}(n) + \\
& z_d(n)(t_{EH}(n) + t_{BC}(n)) + z_e(n)(t_{EH}(n) + t_{IT}(n)) + \\
& z_f(n)(t_{BC}(n) + t_{IT}(n)) + z_g(n)(t_{EH}(n) + t_{BC}(n) + t_{IT}(n)) = 1
\end{aligned} \tag{1}
$$

令 $E_{H,m}(n)$ 、$E_{T,m}(n)$ 分别表示在时隙 n 源端的混合发射器以 m 模式运行时收集的能量和消耗的能量，消耗的能量包括 BC 模式下电路消耗的能量 $P_{c,BC}t_{BC}(n)$ 、IT 模式下电路消耗的能量 $P_{c,IT}t_{IT}(n)$ 、发送状态更新包消耗的能量。因此，对于源端收集的能量 $E_{H,m}(n)$ 和消耗的能量 $E_{T,m}(n)$，可以分别表示为

$$
E_{H,m}(n) = \begin{cases}
\eta P h(n) t_{EH}(n) & \text{if } m \in \{a,e\} \\
(1-\alpha(n))\eta P h(n) t_{BC}(n) & \text{if } m \in \{b,f\} \\
0 & \text{if } m = c \\
\eta P h(n)(t_{EH}(n) + (1-\alpha(n)) t_{BC}(n)) & \text{if } m \in \{d,g\}
\end{cases} \tag{2}
$$

其中 $\eta \in (0,1)$ 为 RF 到 DC 的能量转换效率， $\alpha(n) \in [0,1]$ 表示 n 时隙源端的反向散射系数；

$$
E_{T,m}(n) = \begin{cases}
0 & \text{if } m = a \\
P_{c,BC} t_{BC}(n) & \text{if } m \in \{b,d\} \\
P_{c,IT} t_{IT}(n) + p(n) t_{IT}(n) & \text{if } m \in \{c,e\} \\
P_{c,BC} t_{BC}(n) + P_{c,IT} t_{IT}(n) + p(n) t_{IT}(n) & \text{if } m \in \{f,g\}
\end{cases} \tag{3}
$$

其中， $p(n)$ 表示 n 时隙源端主动信息传输的发射功率。根据香农公式，则 n 时隙内 BC 模式下传输的数据包大小为

$$R_{BC}=t_{BC}(n)\log_2\left(1+\frac{\alpha(n)Ph(n)g(n)}{\delta^2}\right) \quad (4)$$

n 时隙内 IT 模式下传输的数据包大小为

$$R_{IT}=t_{IT}(n)\log_2\left(1+\frac{p(n)g(n)}{\delta^2}\right) \quad (5)$$

若源端在 n 时隙决定传输 M 比特的状态更新包，则反向散射系数 $\alpha(n)$ 和主动信息发射功率 $p(n)$ 需满足如下约束：

$$R_{BC}+R_{IT}\geq M \quad (6)$$

令电池能量的最大量化级别表示为 $b_{\max}$，用 $B(n)\in\{0,e_q,2e_q,...,B_{\max}\}$ 表示 n 时隙源端的电池能量状态，其中 $e_q=\frac{B_{\max}}{b_{\max}}$ 表示能量量子。$B(n)$ 满足如下能量因果关系：

$$B(n)-w(n)E_s+E_{H,m}(n)-E_{T,m}(n)\geq 0 \quad (7)$$

因此，源端的电池能量 $B(n)$ 的变化可以表示为

$$B(n+1)=\min\{\max\{B(n)-w(n)E_s+E_{H,m}(n)-E_{T,m}(n),0\},B_{\max}\} \quad (8)$$

1.3 信息年龄模型

AoI 定义为自源端生成的最新更新到达目的地所经历的时间。令 $C(n)\in\{1,2,...,C_{\max}\}$ 和 $A(n)\in\{1,2,...,A_{\max}\}$ 分别表示 n 时隙源端的 AoI 和目的端的 AoI，$C_{\max}$ 和 $A_{\max}$ 分别表示源端和目的端的 AoI 上界。假设源端执行状态采样需要花费 1 个时隙的时间成本以及 E_s 大小的能量成本。若源端决定进行状态采样，则由于 1 个时隙的采样时间成本，$C(n)$ 保持不变，$C(n+1)$ 重置为 1，否则 $C(n+1)$ 线性增加 1。因此，源端 AoI 的动态变化可以表示为

$$C(n+1)=\begin{cases}1 & \text{if } a(n)=(1,z_m(n))\\ \min\{C_{\max},C(n)+1\} & \text{if } a(n)=(0,z_m(n))\end{cases} \quad (9)$$

其中，$m\in\mathcal{M}$。为了简化表示，上述等式可以重写为

$$C(n+1)=(1-w(n))\min\{C_{\max},C(n)+1\}+w(n) \quad (10)$$

同时，假设源端传输状态更新需要 1 个时隙的传输时间。若源端决定进行状态更新，则 $A(n)$ 重置为 $C(n)+1$，否则 $A(n)$ 线性增加 1。因此，$A(n)$ 的动态变化可以表示为

$$A(n+1)=\begin{cases}\min\{A_{\max},C(n)+1\} & \text{if } a(n)=(w(n),z_k(n))\\ \min\{A_{\max},A(n)+1\} & \text{if } a(n)=(w(n),z_a(n))\end{cases} \quad (11)$$

其中，$k\in\mathcal{M}'$。为了简化表示，$A(n+1)$ 可以通过以下约束表示：

$$A(n+1)=z_k(n)\min\{A_{\max},C(n)+1\}+z_a(n)\min\{A_{\max},A(n)+1\} \quad (12)$$

1.4 优化问题

令 $\pi=\{x(0),x(1),...,x(N)\}\in\Pi$ 表示源端采取的一个确定性决策，它决定了源端每个时隙的状态采样和更新模式决策。其中 $x(n)$ 为 n 时隙源端采取的某个状态采样动作和更新模式，Π 为所有可能的策略集合。若源端采取策略 π，则目的端的长期平均 AoI 可以表示为

$$\overline{A}^{\pi}\triangleq\limsup_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}\mathbb{E}_{\pi}[A(n)], \quad (13)$$

本文的目标是通过寻找年龄最优策略 π^* 来最小化目的端的长期平均 AoI。因此，寻找年龄最优策略 π^* 对应于求解以下问题(P1)：

$$\text{(P1):}\quad \min_{\pi\in\Pi}\ \limsup_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}\mathbb{E}_{\pi}[A(n)]$$

$$\text{s.t.}\quad t_{EH}(n)\in[0,1],t_{BC}(n)\in[0,1],t_{IT}(n)\in[0,1] \quad (14)$$

$$w(n)\in\{0,1\} \quad (15)$$

$$z_m(n)\in\{0,1\},\ \sum_{m\in\mathcal{M}}z_m(n)=1 \quad (16)$$

$$(1),(6),(8),(10),(12)$$

2 最优决策策略

信道状态随时间的独立性导致了源端的能量状态及其能量状态转换的不确定性，因此最小化长期平均 AoI 问题是一个随机优化问题。为了求解这个问题，本文首先将其转换为 MDP 问题，然后针对环境动态信息已知的情况，在 2.3 节使用基于模型的相关值迭代算法对问题进行求解；针对环境动态信息未知的情况，在 2.4 节提出了一个无模型的 Q 学习算法求解问题。

2.1 马尔可夫决策过程

由于信道增益 $h(n)$、$g(n)$ 随时间变化的独立性以及源端的电池能量 $B(n)$、源端和目的端的信息年龄 $C(n)$、$A(n)$ 动态变化过程的马尔可夫性，因此可以将最小化长期平均 AoI 问题建模为无限时域的 MDP 问题。根据[20]，下面对 MDP 的主要组成成分进行详细的介绍。

a) 状态空间：由于实际的信道增益是连续随机变量，因此本文采用 FSMC 模型[21]，将信道增益等概率划分为 κ 个离散信道增益。在这种情况下，可以定义 n 时隙的系统状态为 $s(n)\triangleq\{B(n),A(n),C(n),h(n),g(n)\}\in S$，其中，$S$ 是包含所有可能系统状态的状态空间，它是一个有限集合。

b) 动作空间：在 n 时隙，源端需要决定传感器的采样动作 $w(n)$ 和混合发射器的更新模式 $z_m(n)$，同时确定更新模式的运行参数(包括反向散射系数 $\alpha(n)$、主动信息传输功率 $p(n)$、模式时间分配向量 $t(n)$)。因此，在 $s(n)$ 状态下源端采取的动作可以表示为：$x(s(n))\triangleq\{w(n),z_m(n),\alpha(n),p(n),t(n)\}\in\chi(s)$。其中，$\chi(s)$ 表示系统状态 $s(n)$ 下的动作空间。

c) 转移概率：为了简化表示，使用 $s=\{B,A,C,h,g\}$ 表示当前时隙的系统状态，$s'=\{B',A',C',h',g'\}$ 表示下一时隙的系统状态。由于状态变量之间相互独立，因此在给定当前的系统状态 s 和采取的动作 $x(s)$ 下，从 s 转移到 s' 的概率为

$$\begin{aligned}&\mathbb{P}(s'|s,x(s))\triangleq\mathbb{P}(B',A',C',h',g'|s,x(s))=\\&\mathbb{P}(B'|B,h,g,x(s))\mathbb{P}(A'|A,C,x(s))\mathbb{P}(C'|C,x(s))\mathbb{P}(h')\mathbb{P}(g')\end{aligned} \quad (17)$$

d) 奖励函数：令 $G(s,x(s))$ 表示在 n 时隙，系统状态 s 下采取动作 $x(s)$ 的即时成本，则 $G(s,x(s))$ 可以定义为

$$G(s,x(s))=A' \quad (18)$$

2.2 问题转换

根据 2.1 节对 MDP 组成成分的表述，优化问题(P1)的系统状态空间和动作空间是有限的，它可以转换为一个有限状态有限动作的平均成本 MDP 问题。特别地，优化问题的每阶段平均成本对应 MDP 问题的奖励函数(18)。因此，在给定初始状态 s_0 的情况下，可以重写问题(P1)为

$$\text{(P2):}\quad \min_{\pi\in\Pi}\ \limsup_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}\mathbb{E}_{\pi}[G(s,x(s))|s_0]$$

$$\text{s.t.}\quad (1),(6),(8),(10),(12),(14)\sim(16)$$

2.3 相关值迭代算法

若对于任意时隙 n_1、n_2，当 $s(n_1)=s(n_2)$ 时，如果有 $x(n_1)=x(n_2)$，则称策略 $\pi\in\Pi$ 是确定性平稳策略，其中 Π 为所有可能的确定性平稳策略集合。由于问题(P2)为有限状态有限动作的 MDP 问题，所以存在一个最优的确定性平稳策略[20]。此外，由于策略是平稳的，因此在下文中时间索引可以被忽略。根据[22]，对于平均成本 MDP 问题，可以通过求解以下贝尔曼方程获取最优策略 π^*：

$$\overline{A}^*+V(s)=\min_{x\in\chi(s)}Q(s,x),\forall s\in S, \quad (19)$$

其中，$\overline{A}^*$ 为最优长期平均 AoI，$V(s)$ 是相关值函数，定义为

$$V(s)=\min_{x\in\chi(s)}(Q(s,x)-Q(s_0,x_0)) \quad (20)$$

s_0 可以是任意给定的初始状态，动作值函数 $Q(s,x)$ 定义为

$$Q(s,x)=G(s,x)+\sum_{s'\in S}\mathbb{P}(s'|s,x)V(s') \quad (21)$$

因此，可以通过求解下式获得长期平均 AoI 最优策略 π^*：

$$\pi^*=\arg\min_{x\in\chi(s)}Q(s,x). \quad (22)$$

为了获得 $\bar{A}^*$ 和 π^*，在已知信道转移概率的情况下，本文采用相关值迭代算法(relative value iteration algorithm, RVIA)[22]迭代地求解贝尔曼方程(19)。特别地，对于任意初始状态 s_0，在 RVIA 的第 $k+1$ 次迭代中，有如下等式：

$$Q(s,x)^{(k+1)}=G(s,x)+\sum_{s'\in S}\mathbb{P}(s'|s,x)V(s')^{(k)} \tag{23}$$

$$A^{(k+1)}=\min_{x\in\chi(s)}Q(s,x)^{(k+1)} \tag{24}$$

$$V(s)^{(k+1)}=\min_{x\in\chi(s)}\left(Q(s,x)^{(k+1)}-Q(s_0,x_0)^{(k+1)}\right) \tag{25}$$

令 $c_{\max}^{k+1}-c_{\min}^{k+1}$ 表示第 $k+1$ 次迭代的贝尔曼误差，其中 $c_{\max}^{k+1}$ 和 $c_{\min}^{k+1}$ 分别定义为

$$c_{\max}^{k+1}=\max_{s\in S}\left|V(s)^{(k+1)}-V(s)^{(k)}\right| \tag{26}$$

$$c_{\min}^{k+1}=\min_{s\in S}\left|V(s)^{(k+1)}-V(s)^{(k)}\right| \tag{27}$$

当第 k 次迭代的贝尔曼误差满足 $|c_{\max}^{k}-c_{\min}^{k}|\le\epsilon$ 时，$\bar{A}^{(k)}$ 将收敛到每阶段最优平均成本 $\bar{A}^*$，此时通过(22)式即可获得对应的最优策略 π^*。算法的具体步骤如算法 1 所示。

算法 1　相关值迭代算法

输入：初始系统状态 s_0，以及贝尔曼误差阈值 ϵ。

输出：$\bar{A}^*$，以及最优策略 π^*。

a) 初始化 $k=0$，$V(s)^{(0)}=0$ 以及 $|c_{\max}^{0}-c_{\min}^{0}|>\epsilon$。

b) 当 $|c_{\max}^{k}-c_{\min}^{k}|>\epsilon$ 时，重复执行以下步骤；

c) 计算每个状态 $s\in S$ 的 $Q(s,x)^{(k+1)}$ 以及 $\bar{A}^{(k+1)}$；

d) 令 $V(s)^{(k)}=V(s)^{(k+1)}$，计算 $V(s)^{(k+1)}=\min_{x\in\chi(s)}\left(Q(s,x)^{(k+1)}-Q(s_0,x_0)^{(k+1)}\right)$，以及 $c_{\max}^{k+1}$ 和 $c_{\min}^{k+1}$，更新迭代步数 $k=k+1$ 后转步骤 b)。

e) 通过计算(22)式可以得到最优策略 π^*。

2.4　Q 学习算法

在实际环境中，信道状态的转移概率通常是难以获得的，因此本文采用一种无模型的 Q 学习在线算法[18]求解问题(P2)，迭代地寻找最优策略。具体的来说，在 Q 学习的算法过程中，源端通过不断地与环境进行试错交互，估计和学习最优的动作值函数；然后源端将根据学习到的 Q 值选择当前状态下的动作。为了保证估计的动作值函数最终能够收敛到最优动作值函数，本文使用 ε 贪婪策略来权衡探索和利用，它能保证探索到足够丰富的环境状态，同时能利用探索到的状态信息来最小化系统的长期平均 AoI。因此，在每个时隙中，源端将以 ε 的概率选择随机动作，以 $1-\varepsilon$ 的概率选择最优动作。在数学上，遵循 ε 贪婪策略的动作选择可以表示为

$$x(n)=\begin{cases}\arg\min_{x\in\chi(s)}Q(s(n),x(n)) & \text{if } \varepsilon<p_r\le 1\\ x_{rd}\in\chi(s) & \text{if } p_r\le\varepsilon\end{cases} \tag{28}$$

其中，$p_r\sim u(0,1)$ 为当前时隙下随机生成的概率，x_{rd} 表示随机选择的动作。特别地，在给定状态动作对 (s,x) 下，n 时隙处 Q 学习的迭代更新公式可以表示如下：

$$\begin{aligned}&Q(s(n),x(n))=(1-\gamma(n))Q(s(n),x(n))+\gamma(n)(G(s(n),x(n))+\\&\min_{x\in\chi(s)}Q(s(n+1),x(n))-\min_{x_0\in\chi(s_0)}Q(s_0,x_0))\end{aligned} \tag{29}$$

其中，$\gamma(n)$ 表示时隙 n 处的学习率。为了加快 Q 学习算法的学习速度并且保证源端探索到足够的状态信息，通常需要在迭代的初始时期设置较大的学习率 $\gamma(n)$ 和贪婪率 ε。另一方面，随着迭代次数的增加，需要逐渐减少学习率和贪婪率，以便估计的动作值函数可以快速平稳地收敛到最优动作值函数。Q 学习算法的详细步骤如算法 2 所示。

算法 2　Q 学习算法

输入：初始系统状态 s_0，学习率 $\gamma(n)$ 和贪婪率 ε。

输出：学习到的策略 π^*。

a) 初始化 $n=0$，$Q(s,x)=0,\forall s\in S,x\in\chi(s)$ 以及学习率 $\gamma(n)$ 和贪婪率 ε，随机选择一个初始状态 s_0。

b) 当时隙 n 小于预设值时，重复执行以下步骤；

c) 在当前状态 $s(n)$ 下根据 ε 贪婪策略选择动作 $x(n)$；以 ε 概率选择随机动作，以 $1-\varepsilon$ 概率选择最优动作。

d) 采取动作 $x(n)$，与环境交互获得环境回报 $G(s(n),x(n))$ 和下一个系统状态 $s(n+1)$。

e) 通过计算(29)式更新动作值 $Q(s(n),x(n))$，在更新时隙数 $n=n+1$ 后转步骤 b)。

f) 最后计算 $\pi^*=\arg\min_{x\in\chi(s)}Q(s,x)$ 得到学习到的策略 π^*。

3　仿真结果及性能分析

在这一部分中，本文对所提的联合采样和混合反向散射通信更新策略的性能进行了分析。为了评估所提策略的性能，本文与联合采样和 WPC 更新策略(表示为 A 策略)[12]、联合采样和 BC 更新策略(表示为 B 策略)进行了对比。仿真结果展示了在信道动态信息已知的情况下算法 1 的性能，以及在缺乏信道动态信息的情况下算法 2 提出的 Q 学习算法的性能。

3.1　仿真参数设置

在仿真中，设置源端的能量收集效率 $\eta=0.7$，目的端的噪声功率 $\delta^2=-95\,\text{dBm}$ [12]。能量发射器 ET 到源端 S 的距离 d_{ES} 以及源端 S 到目的端 D 的距离 d_{SD} 为 $10\,\text{m}$。路径损失建模为 $L=20+20\log_{10}d$ [17,23]，其中 d 是信道链路距离。设置每个时隙的持续时间为 1 秒，能量发射器的发射功率 P 为 $25\,\text{dBm}$，源端电池容量为 $B_{\max}=10\eta P\bar{h}$ [17]，其中 $\bar{h}$ 为源端上行链路的平均信道增益。状态采样的能量成本 $E_s=3e_q$，反向散射通信和主动信息传输的电路能耗分别设置为 $P_{c,BC}=8.9\,\mu\text{W}$，$P_{c,IT}=113\,\mu\text{W}$ [17,24]。源端的反向散射系数 $\alpha(n)$ 被离散化为 5 级，其余状态和动作变量被离散化为 10 级。特别地，由于采用等概率的方法划分信道增益，因此信道状态转移概率为 $\mathbb{P}(h')=\mathbb{P}(g')=1/K=0.1$。

3.2　性能分析

仿真结果图 3~图 5 展示了在信道动态信息已知情况下相关值迭代算法的性能。其中，图 3 显示了 ET 的发射功率变化时不同策略的可实现最优长期平均 AoI，更新包的大小设定为 $M=18\,\text{Mbits}$。可以看到，无论 ET 的发射功率如何变化，本文提出的策略明显优于联合采样和 WPC 更新策略以及联合采样和 BC 更新策略。这是由于所提策略结合了 BC 模式低功耗的特点和主动 IT 模式高速率的特点，可以在不同信道状态下选择最优的更新包传输模式。具体地，在所提策略下，ET 的发射功率较小时，源端电池中存储的能量较少，它可以选择 BC 模式或者 BC-IT 等组合模式进行更新包的紧急传输。ET 的发射功率较大时，源端可以存储较多的能量在电池中，因此它将有更多的机会在信道条件差的情况下，将更新包发送到目的地。

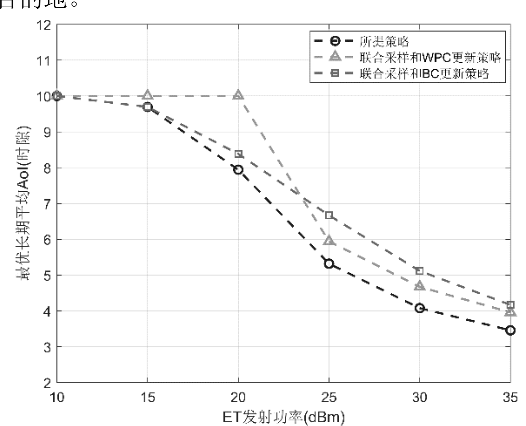


图 3　最优长期平均 AoI 对比能量发射器的功率

Fig. 3　Optimal long term average aoi versus power of energy transmitter

此外，还可以观察到，在 ET 的发射功率较低时，B 策略的可实现 AoI 低于 A 策略，而在 ET 的发射功率较高时，A 策略的可实现平均 AoI 低于 B 策略。这是由于 A 策略所需要的更新能量成本较高，在 ET 的发射功率较低时，源端没有足够的能量及时地进行更新的传输，导致可实现的最优平均 AoI 要比采用 B 策略的高。但是，随着 ET 发射功率的增加，源端收集的能量也逐渐增加，由于主动 IT 模式相比 BC 模式传输速率更高的特点，使得 A 策略的可实现最优平均 AoI 逐渐低于 B 策略。

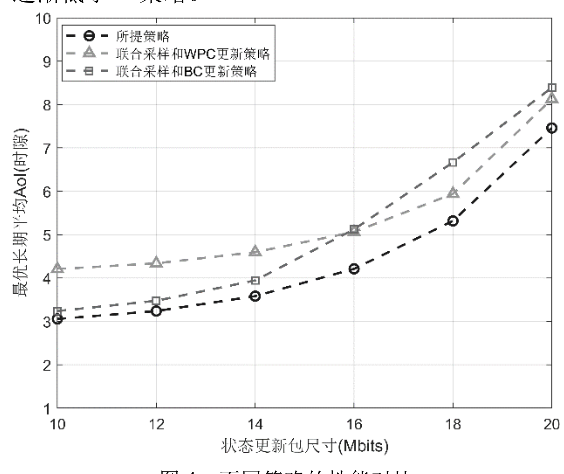


图 4 不同策略的性能对比

Fig. 4 Performance comparisons of different policy

在图 4 中，比较了当更新包大小 M 变化时，不同策略的可实现最优长期平均 AoI 变化。本文所提策略的性能要优于 A 策略和 B 策略，并且随着状态更新包尺寸的增加，所有策略的最优平均 AoI 都单调增加。还可观察到，在更新包较小时，B 策略的平均 AoI 性能明显优于 A 策略；然而，当更新包较大时，A 策略的平均 AoI 性能要优于 B 策略，这是因为相比 BC 模式，主动 IT 模式的传输速率更快，可以传输更大的更新包。

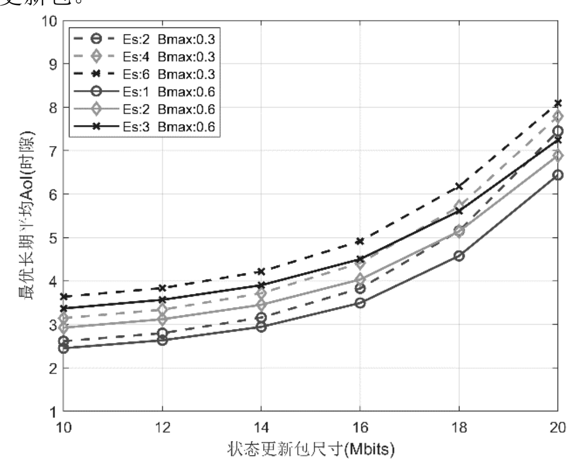


图 5 最优长期平均 AoI 对比更新包大小

Fig. 5 Optimal long term average aoi versus state update packet size

图 5 绘制了对于不同的采样成本 E_s 和电池容量 B_{max}，最优长期平均 AoI 对比更新包尺寸的情况。特别地，由于在参数设置 $B_{max}=0.6\text{mJ}$ 情况下的单位能量量子是参数设置 $B_{max}=0.3\text{mJ}$ 情况下的两倍，因此为了保证在对照组的电池容量变化时，对应的采样能量成本相等，需要分别设置当 $B_{max}=0.6\text{mJ}$ 时的采样成本为 $E_s=1e_q$ 、$E_s=2e_q$ 和 $E_s=3e_q$ 。从仿真结果中可以明显看出，随着 E_s 的减小或者 B_{max} 的增大，系统的最优长期平均 AoI 减小。这是因为 E_s 越小，源端就能节省越多能量；B_{max} 越大，源端就能存储越多的能量，这都增加了源端在未来持续运行的可能性。同时，由于增大电池容量将能传输更大的状态更新包，因此在更新包较大时增大电池容量相比减少采样能量成本更能提升系统的 AoI 性能；并且随着状态更新包尺寸的增加，这一性能提升差异变得越来越明显。

图 6 展示了基于模型的相关值迭代算法和无模型 Q 学习算法在收敛后 10^4 时隙中得出的系统平均 AoI 性能。特别地，由于相关值迭代算法知道环境的精确统计模型(如信道状态转移概率等)，因此它作为 Q 学习算法的性能下界(最优性能)。可以观察到，两种算法的平均 AoI 都随着 ET 发射功率的增加而下降，并且 Q 学习算法的性能非常接近相关值迭代算法的性能。具体而言，Q 学习算法的性能在整体上接近相关值迭代算法性能的 96.23%。因此，即使源端在缺乏信道动态信息的情况下，采用 Q 学习算法依然可以达到较高的系统 AoI 性能。

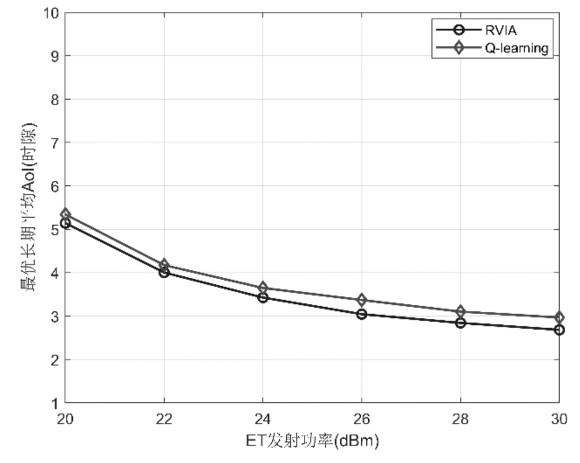


图 6 提出的 Q 学习算法的性能分析

Fig. 6 Performance analysis of the proposed Q-learning algorithm

4 结束语

本文研究了一种反向散射辅助无线供电通信系统的长期平均 AoI 最小化问题。为了提高系统的 AoI 性能，提出了联合采样和 HBC 更新策略，其中源端可以动态地选择传感器的采样动作和发射器的更新模式。为了获得最优策略，首先将问题建模为一个有限状态和有限动作的无限时域平均成本 MDP 问题，然后在信道动态信息已知的场景下，通过相关值迭代算法对问题进行迭代求解；在信道动态信息未知的场景下，采用无模型的 Q 学习算法学习最优策略。最终，数值结果表明，本文提出的策略明显优于联合采样和 WPC 更新策略、联合采样和 BC 更新策略;同时发现，采用 Q 学习算法可以在缺乏信道动态信息的情况下，通过试错交互和学习也可以达到较高的 AoI 性能。在未来的工作中，将考虑一个反向散射辅助无线供能通信的多源双跳中继网络的场景，通过深度强化学习算法寻找年龄最优策略，以优化系统的 AoI 性能。

参考文献：

[1] Abd-Elmagid M A, Pappas N, Dhillon H S. On the role of age of information in the internet of things [J]. IEEE Communications Magazine, 2019, 57 (12): 72-77.

[2] Kaul S, Yates R, Gruteser M. Real-time status: how often should one update? [C]// Proc of IEEE INFOCOM. Piscataway, NJ: IEEE Press, 2012: 2731-2735.

[3] Ma D, Lan G, Hassan M, *et al.* Sensing, computing, and communications for energy harvesting IoTs: a survey [J]. IEEE Communications Surveys & Tutorials, 2020, 22 (2): 1222-1250.

[4] Ponnimbaduge Perera T D, Jayakody D N K, Sharma S K, *et al.* Simultaneous wireless information and power transfer (SWIPT): recent advances and future challenges [J]. IEEE Communications Surveys & Tutorials, 2018, 20 (1): 264-302.

[5] 孙径舟, 王乐涵, 孙宇璇, 等. 面向 6G 网络的信息时效性度量及研

究进展 [J]. 电信科学, 2021, 37 (6): 3-13. (Sun Jingzhou, Wang Lehan, Sun Yuxuan, *et al.* Information timeliness metrics and research progress for 6G network [J]. Telecommunications Science, 2021, 37 (6): 3-13.)

[6] Ponnimbaduge Perera T D, Jayakody D N K, Pitas I, *et al.* Age of information in SWIPT-enabled wireless communication system for 5GB [J]. IEEE Wireless Communications, 2020, 27 (5): 162-167.

[7] Arafa A, Yang Jing, Ulukus S, *et al.* Age-minimal transmission for energy harvesting sensors with finite batteries: online policies [J]. IEEE Trans on Information Theory, 2020, 66 (1): 534-556.

[8] Leng Shiyang, Yener A. Age of information minimization for an energy harvesting cognitive radio [J]. IEEE Trans on Cognitive Communications and Networking, 2019, 5 (2): 427-439.

[9] Krikidis I. Average age of information in wireless powered sensor networks [J]. IEEE Communications Letters, 2019, 8 (2): 628-631.

[10] Abd-Elmagid M A, Dhillon H S, Pappas N. A reinforcement learning framework for optimizing age of information in RF-powered communication systems [J]. IEEE Trans on Communications, 2020, 68 (8): 4747-4760.

[11] 刘玲珊，熊轲，张煜，等. 信息年龄受限下最小化无人机辅助无线供能网络的能耗：一种基于 DQN 的方法 [J]. 南京大学学报：自然科学, 2021, 57 (5): 847-856. (Liu Lingshan, Xiong Ke, Zhang Yu, *et al.* Energy minimization in UAV-assisted wireless powered sensor networks with AoI constraints: A DQN-based approach [J]. Journal of Nanjing University: Natural Science, 2021, 57 (5): 847-856.)

[12] Abd-Elmagid M A, Dhillon H S, Pappas N. AoI-optimal joint sampling and updating for wireless powered communication systems [J]. IEEE Trans on Vehicular Technology, 2020, 69 (11): 14110-14115.

[13] Liu V, Parks A, Talla V, *et al.* Ambient backscatter: wireless communication out of thin air [J]. ACM SIGCOMM Computer Communication Review, 2013, 43 (4): 39-50.

[14] Lu Xiao, Niyato D, Jiang Hai, *et al.* Ambient backscatter assisted wireless powered communications [J]. IEEE Wireless Communications, 2018, 25 (2): 170-177.

[15] Li Dong, Peng Wei, Liang Yingchang. Hybrid ambient backscatter communication systems with harvest-then-transmit protocols [J]. IEEE Access, 2018, 6: 45288-45298.

[16] 叶迎晖，施丽琴，卢光跃. 反向散射辅助的无线供能通信网络中用户能效公平性研究 [J]. 通信学报, 2020, 41 (7): 84-94. (Ye Yinghui, Shi Liqin, Lu Guangyue. User-centric energy efficiency fairness in backscatter-assisted wireless powered communication network [J]. Journal on Communications, 2020, 41 (7): 84-94.)

[17] Long Yusi, Huang Gaofei, Tang Dong, *et al.* Achieving high throughput in wireless networks with hybrid backscatter and wireless-powered communications [J]. IEEE Internet of Things Journal, 2021, 8 (13): 10896-10910.

[18] Sutton R S, Barto A G. Reinforcement Learning: An Introduction [M]. Cambridge, MA: MIT Press, 2018.

[19] Zhou Bo, Saad W. Joint status sampling and updating for minimizing age of information in the internet of things [J]. IEEE Trans on Communications, 2019, 67 (11): 7468-7482.

[20] Puterman M L. Markov decision processes: discrete stochastic dynamic programming [M]. New York: Wiley, 1994.

[21] Sadeghi P, Kennedy R A, Rapajic P B, *et al.* Finite-state Markov modeling of fading channels-a survey of principles and applications [J]. IEEE Signal Processing Magazine, 2008, 25 (5): 57-80.

[22] Bertsekas D P. Dynamic programming and optimal control [M]. Belmont, MA: Athena Scientific, 2005.

[23] Zhou Xun, Zhang Rui, Ho C K. Wireless information and power transfer: architecture design and rate-energy tradeoff [J]. IEEE Trans on Communications, 2013, 61 (11): 4754-4767.

[24] Lu Xiao, Jiang Hai, Niyato D, *et al.* Wireless powered device to device communications with ambient backscattering: performance modeling and analysis [J]. IEEE Trans on Wireless Communications, 2018, 17 (3): 1528-1544.